ISM 6562

Big Data for Business

Dr. Tim Smith

UNIVERSITY of SOUTH FLORIDA

Guiding Agenda

- 10:00-10:25 - Introductions
- 10:25-10:40 - Review syllabus
- 10:40-11:00 – Technical Foundations
    - Git/GitHub
    - GTerminal
- Break
- 11:00-12:00 - Introduction to Analytics, Data Mining, and Big Data
    - Rationality in decision-making
    - The rise of big data
    - Scaling for big data
- 12:00-12:10 - Introduce DataCamp, get everyone signed up
- 12:10-12:20 - Class recap, questions, deliverables this week, and preparing for next class

Meet your professor

- Dr. Tim Smith
  - PhD in Business Information Systems
- Research Interests include technology adoption , organizational routines, health information technology, and application to deep learning models to cryptocurrency and blockchain. Coauthoring a textbook on Business Analytics (Prospect Press).
- 20+ years of professional experience
  - Systems programmer
  - Founder, and president of a software development company
  - Systems Architect
  - Telecom Executive
- In my spare time, I like to travel, hike, fly-fish, and play guitar.

Introduce Yourselves

- Introduce yourself to the class
- Tell them your name, and one interesting thing about yourself:
  - i.e., mine was that I played lead guitar in a heavy metal band

Course Introduction

ISM6562 – Big Data for Business

- Review Syllabus
- Review Academic Integrity Policy (Cheating)
- Review Canvas and how to navigate course content

Server/Dev Environment Introduction

Accessing our class VM

- Let’s walk through connecting to class VM

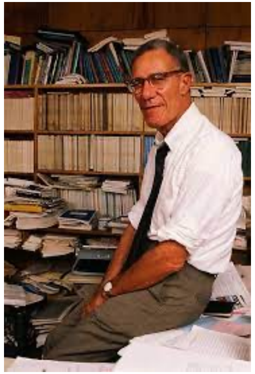
Introduction to Analytics, Data Mining and Big Data

Learning objectives

- Understand what Solow’s Paradox is/was, and how this is changing in today’s data driven world.
- Understand big data, and the 6v’s
- Situate Data Mining and Data Analytics within the large scope of data science.
- Describe the four types of data analytics.
- Describe some examples of the use of data analytics in business.

1987….

“You can see the computer age everywhere but in the productivity statistics“.



Robert Solow, 1987

Today…

Since 2002, the top 5% of companies have increased productivity by 40%, while the other 95% of companies have barely increased productivity at all.

Andy Haldane, Chief Economist for the Bank of England

What explains this?

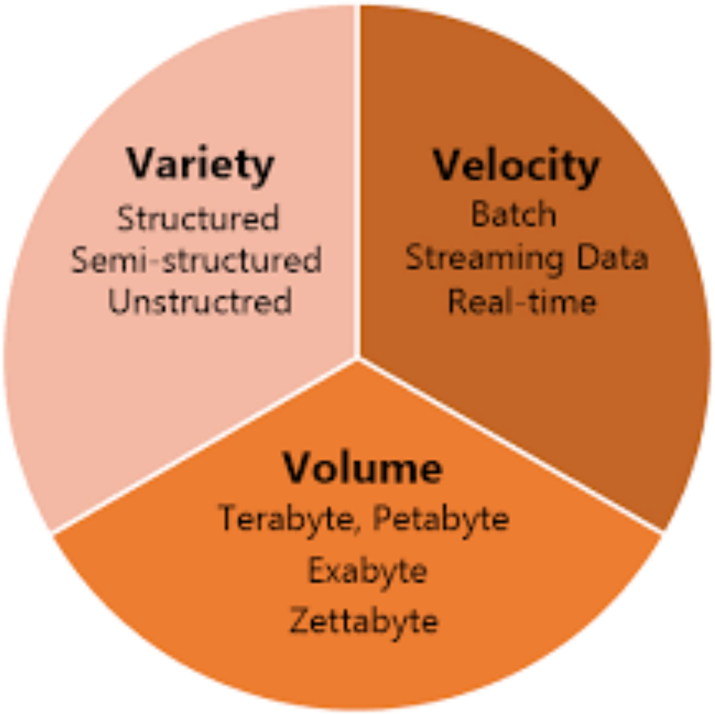
- Some companies have found a solution to Solow’s Paradox and are rapidly improving their productivity.
- Economists continue to explore what’s behind this trend, but a few things are clear
    - The Internet is increasing the rate of consolidation in many industries.
        - Differentiation is even more important today!
    - Effective use of data analytics and machine learning are being leveraged to improve and automate business processes and decision making.
        - Clearly, some companies do a better job at this than others!
        - Data, and access to information from this data is driving tremendous competition advantage and growth.

Data is everywhere…

- 2.5 quintillion bytes of data generated each day
- Over the last two years, we've generated 90% of all the data that has ever been generated!
  - We conduct more than half of our web searches from a mobile phone now.
  - More than 3.7 billion humans use the internet (that's a growth rate of 7.5 percent over 2016).
  - On average, Google now processes more than 40,000 searches EVERY second (3.5 billion searches per day)!

See the infographic here (https://web-assets.domo.com/blog/wp-content/uploads/2017/07/17_domo_data-never-sleeps-5-01.png)

Big Data: The 3vs



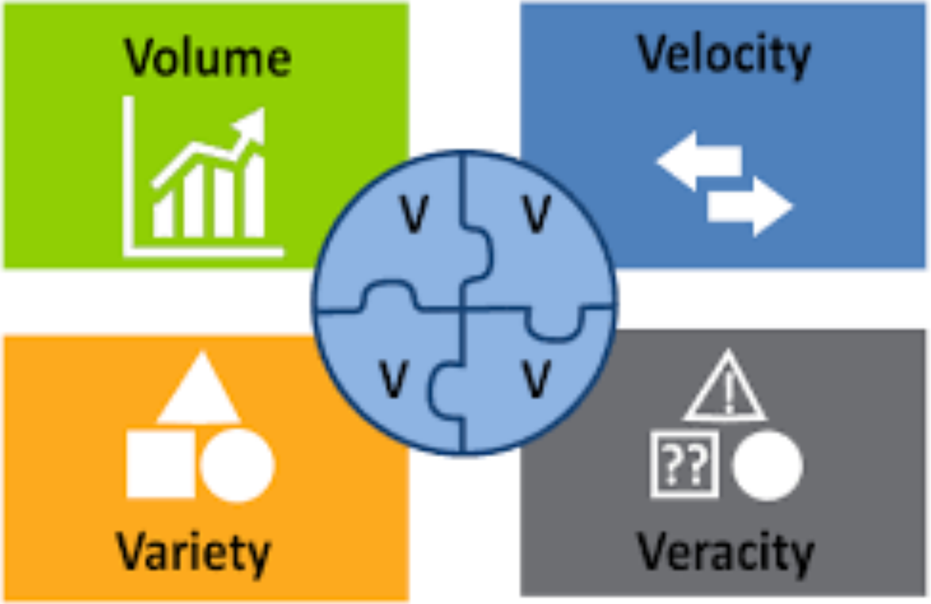
Volume: This refers to the amount of data. With the advent of the internet and the rise of cell phone use and IoT, we are seeing exponential growth in data produced.

Variety: The variety of data is expanding – videos, audio, web interactions, geographic location, images, emails, etc.

Velocity: The speed of data production is increasing.

- In 2022, **333.2 billion emails** are sent every day.
- In 2021, people created **2.5 quintillion bytes** of data every day.
- In 2022, users sent around **650 million Tweets** per day.
- In 2021, people created **2.5 quintillion bytes** of data every day.

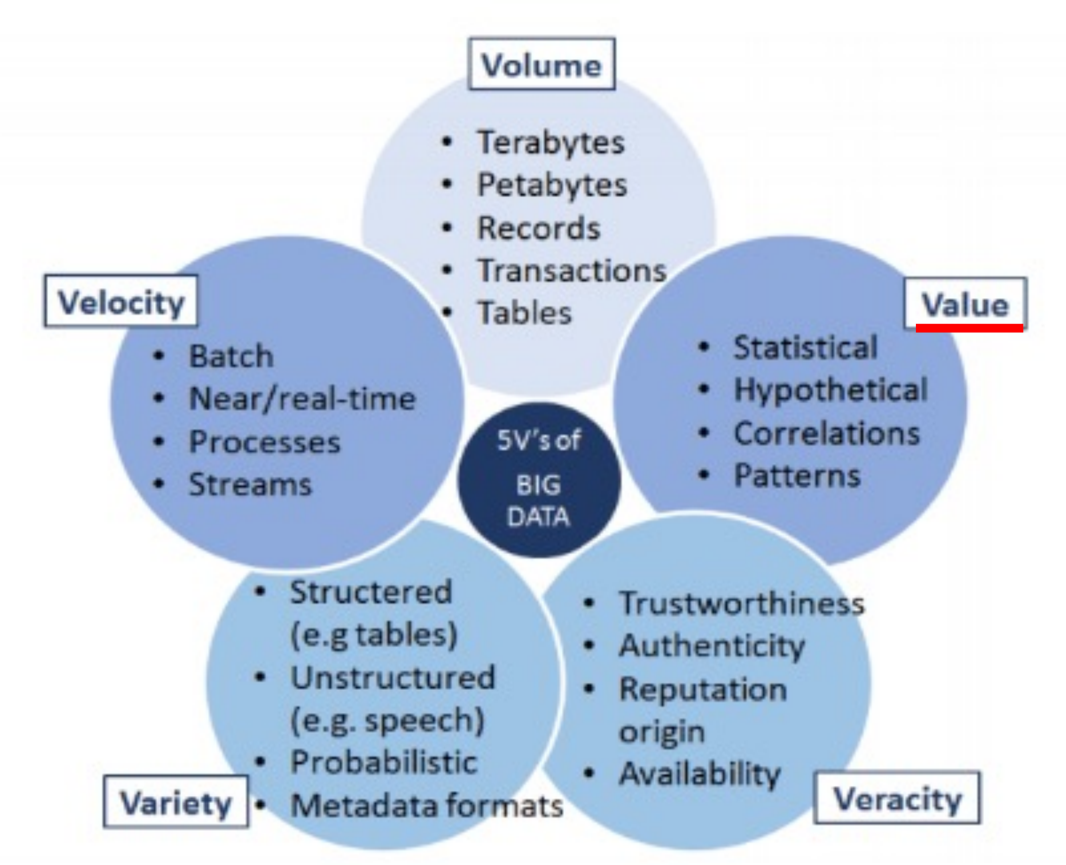
Other Systems: Big Data – 3Vs and 4Vs



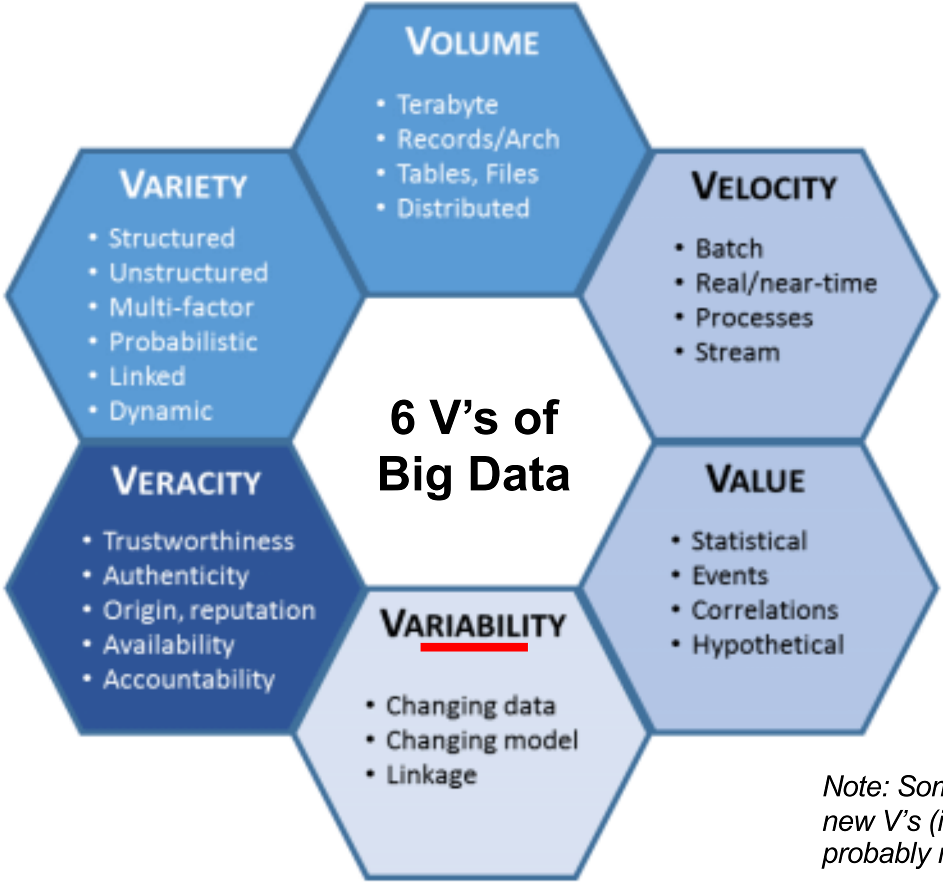
Veracity: Refers to the quality of the data; where the data has been collected from, how it was collected, and how it will be analyzed.

- The veracity of a users data, dictates how reliable and significant the data really is.

Other Systems: Big Data – 5 Vs



Other Systems: Big Data – 6 Vs



Note: Some experts have kept adding new V's (i.e., 17!) – but beyond 6 is probably not necessary or valuable.

History of Big Data

- The term was coined in the 90s
  - typically credited to John Mashey from Silicon Graphics
- But, we have been dealing with ‘big data’ challenges since ancient times:
  - 300 BC: Attempt to capture all existing ‘data’ in the Alexandria library
  - Roman Empire analyzed data to determine optimal distribution of their armies

Big Data Phase 1	**Big Data Phase 2**	**Big Data Phase 3**
Period: 1970-2000	Period: 2000-2010	Period: 2010-present
RDBMS Based structured content: • RDBMS and Data Warehousing	Web-based Unstructured content: • Information retrieval and extraction	Mobile and Sensor Based Content • Entity-centered analysis (i.e., person)

From a business perspective

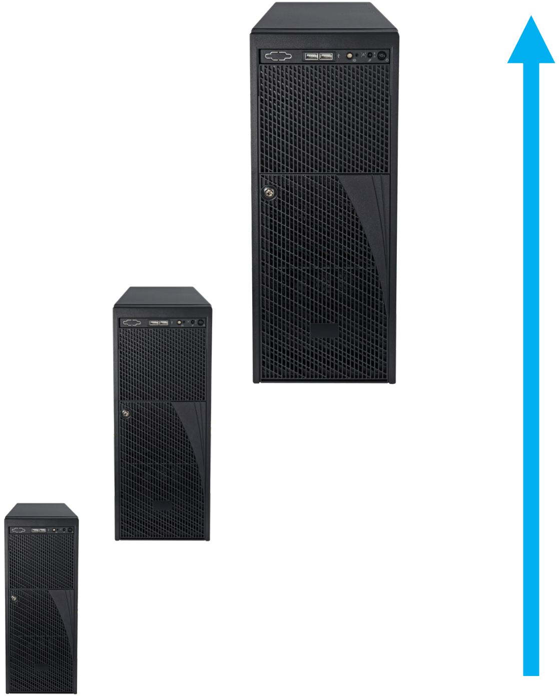
- How do you store and process the data in a specific use-case scenario?

- Will a single computer suffice?
  - Can we utilize a single system, or ‘scale up’ the capacity of a single system to handle the job?

Vertical vs Horizontal Scaling

- Big Data systems make distributing storage and processing across multiple nodes easier to accomplish
  - What was extremely technically challenging and costly to do has become much easier and less costly.

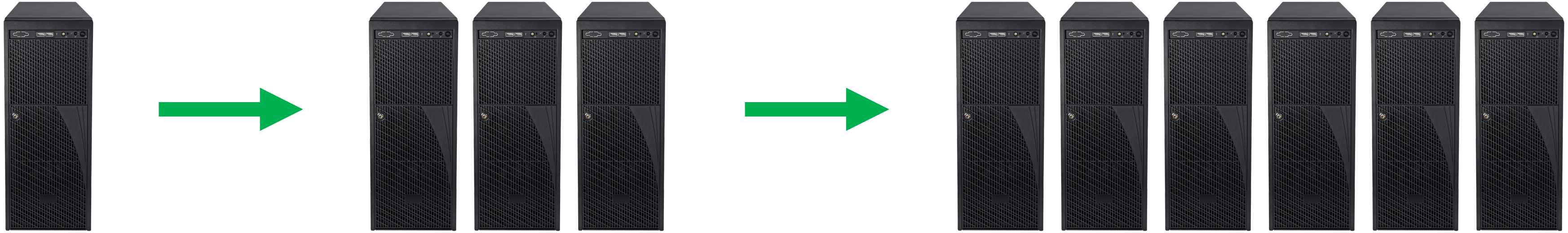
Vertical Scaling



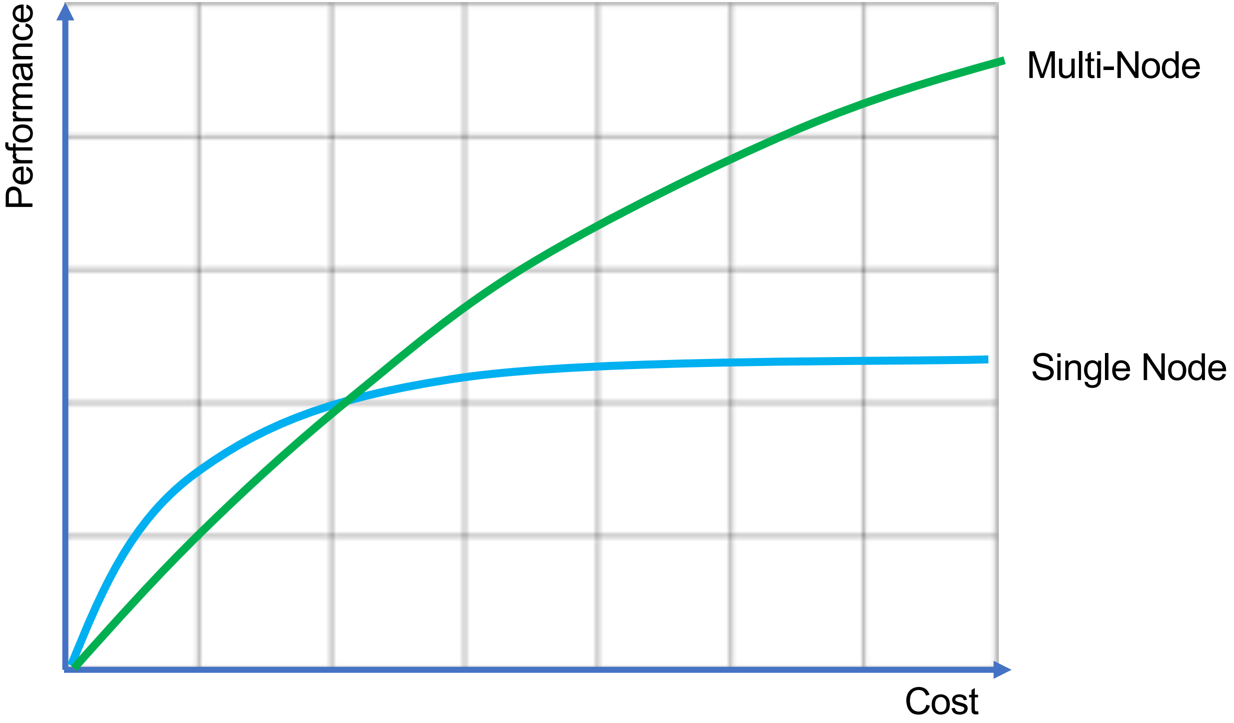
- Increase the number of CPU cores
- Increase the amount of RAM
- Increase the storage disk space

Horizontal Scaling

- Increase the number of systems
- Utilize systems/software technologies that distribute storage and processing across multiple computers



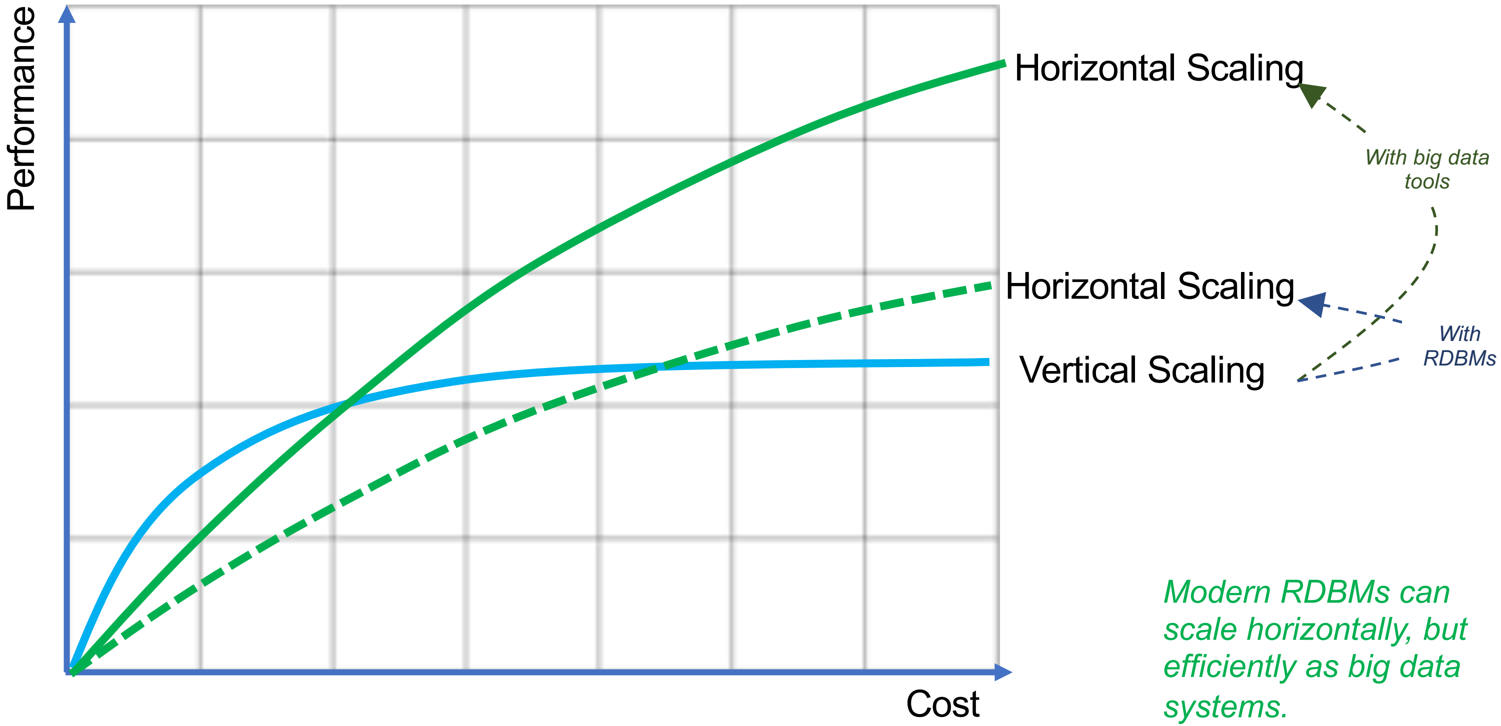
How much performance does the use case require?



Discussion

- What’s an RDBMS?
- What are they good for? What contexts are they possibly not the best solution?

Scaling – Incorporating Big Data Systems



Modern RDBMs can scale horizontally, but efficiently as big data systems.

Data Lake and Data Warehouse

See Notes:

- You should receive an invitation to join or GitHub organization
  - If you don’t already have one, create a github account
    - Use your USF email
  - Accept the invitation to the class organization on GitHub
- ISM6562-GBAIS/ClassMaterial (github.com)

DataCamp

Data Camp

- I’ve arranged for each one of you to have access to DataCamp for five months.
- These courses are part of the supplemental material - I will not assign you any courses.
- Accept the invitation and create an account using you USF email

Summary

Summary

- We have reviewed the syllabus, canvas, and datacamp.
- You are now subscribed to datacamp
- You can connect to the Cyberhub VM Server
- You understand some of the challenges of big data analytics
- You know the 3+ V’s of big data
- You understand the benefits and costs of vertical versus horizontal scaling.

Next Class

- Be sure that you have created a GitHub account.
- Accept the invitation to our Class Organization on GitHub
- We will cover the Technology Foundations of this Course:
  - Introduction to Shell
  - Introduction to Git/GitHub
  - Introduction to VMs and Containers

Happy Learning!

UNIVERSITY of SOUTH FLORIDA